International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 7 Issue 3, May-June 2023 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Study of Digital Media-Based Voice Activity Detection Protocols

Nikhil Kumar[1], Sumit Dalal[2]

[1]Student, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India
[2]Assistant Professor, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

ABSTRACT

Speaker identification is critical for a variety of voice-based applications in safety and surveillance systems, and these kinds of methods are now employed in household appliances for user-controlled device toggling. A comprehensive and language uncertain Voice Activity Detection (VAD) system is critical for Digital Media Content (DMC). VAD systems are utilised for DMC generation in a variety of methods, including supplementing subtitle formation, detecting and correcting subtitle drifting, and sound distortion. The goal of this article is to provide a comprehensive overview of numerous strategies utilised for voice recognition in the entertainment industry. An analysis of several speaker recognition strategies used earlier and those utilised in current studies was explored, and a clear understanding of the superior methodology was discovered through a survey across various literature for more than two decades. We give a comprehensive survey of DNN-based VADs using DMC data concerning accuracy, noise sensitivity, and language agnostic performance in this paper.

KEYWORDS: *Voice Activity Detection, Digital Media Content, Deep Learning*

How to cite this paper: Nikhil Kumar | Sumit Dalal "A Study of Digital Media-Based Voice Activity Detection Protocols" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-7 | Issue-3, June 2023, pp.422-426, URL: www.ijtsrd.com/papers/ijtsrd56376.pdf

IJTSRD56376

INTRODUCTION

In the digital media domain, a computerized scheme that identifies human speech or vocal activity inside an audio segment has a variety of applications [1]. A common VAD for DMC is a difficult problem for several reasons. Initially, DEC voice segments frequently cohabit with other noises such as integrated front or background music, a title track, and several other noises. Relevant background noises such as traffic, shootings, crowd buzz, gate closing, engines, AC, and so on. Next, because DEC is an artistic medium, abnormal speech patterns such as muttering, yelling chanting, and electronic sounds are more common than other conventional speech difficulties. Third, and most significantly, nearly all of past VAD research has concentrated on the English language. Any viable VAD system for DEC must be capable of scaling across numerous languages, locations, and genres. As a result, most previous conventional VAD techniques are no longer readily relevant for DEC-related systems. VAD and its components are shown in Fig 1:

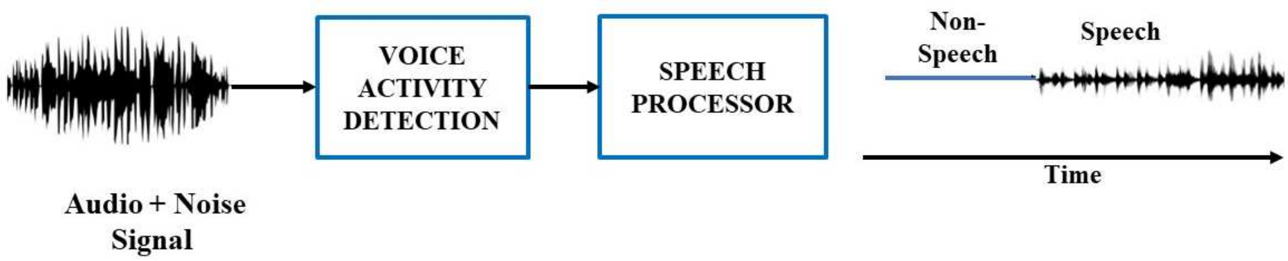

Fig 1: VAD and its Components

RELATED WORK:

For numerous decades, researchers have been working on the creation of VAD systems. To estimate the per-frame likelihood of speech, two GMMs were used, one trained on speech frames and the other on non-speech frames, accompanied by a Hidden Markov Model (HMM) that penalizes modifications between speech and non-

speech states to provide temporal consistency to the anticipation [2]. WebRTC VAD [3] is a GMM-enabled VAD model with input characteristics such as log values of 6 frequency bands ranging from 80 to 4000 Hz. It employs fixed point procedures and is designed for actual time web communication. A better technique would be to train a classifier using time and frequency domain audio data such as mean, variance, higher order moments, coefficient of variation, and percentiles of the distribution. The probability distribution of the audio signal and the MFCCs [4] were used as input to a GMM [5], SVM [6], RF [7], and neural networks(NN) [8]. There are two fundamental issues with these strategies. First, due to the Markov property and tiny discrete state space, models employing HMM cannot learn long-term dependencies. Recent research confirms that using features with a longer length enhances performance because they may represent contextual information more correctly [9]. Secondly, when there is noise in the background with spectral features akin to speech, the efficacy of these approaches degrades [10].

Deep learning for sequencing has made remarkable progress in recent times, particularly for VAD in DEC. Mateju et al. [11] employed a DNN trained on a noise database, combined with output smoothing, to detect speech activity in videos. As the input feature for VAD in movies, Jang et al. [12] used a two-layered DNN with MFCC. Zhang et al. [13] employed a boosted deep neural network DNN to generate several estimates from dissimilar contexts of a single frame using only one DNN and then aggregate the predictions for a superior frame prediction.

ANALYSIS OF DIFFERENT VAD METHODS

A. Gaussian mixture model (GMM) based methods

The authors [14] suggested a GMM model for the log-power distributions of noise and (noisy) speech, in which the spread of both of these factors can be self-adapt in non-stationary conditions. An adaptive threshold based on these two elements' GMM variables indicates a realistic bound between noise and speech, which may result in a precise VAD in a variety of noise circumstances. Some dependability limitations are given to this suggested GMM to increase SHR and NHR. The experimental findings show that the proposed technique performs admirably for SHR and NHR. Authors in [15] have proposed a GMM based VAD model. The Visual speech classifier component was created to assess the likelihood of speech by allocating scores to feature file frames. The feature extractor component generates feature files. They provided a framework for detecting vocal activity using visual articulators. The article specifically evaluated the usefulness of the variations in the visual domain from the speaker's frontal and profile views for the VVAD task. To the best of our knowledge, this is the first attempt at VVAD using profile views. Our investigations indicated that profile perspectives indeed contain essential visual speech knowledge, but, as would be expected, less than upfront data owing to the poor recording of visual data from profile perspectives. Fig. 2 shows a detailed description of GMM based video speech detection method.

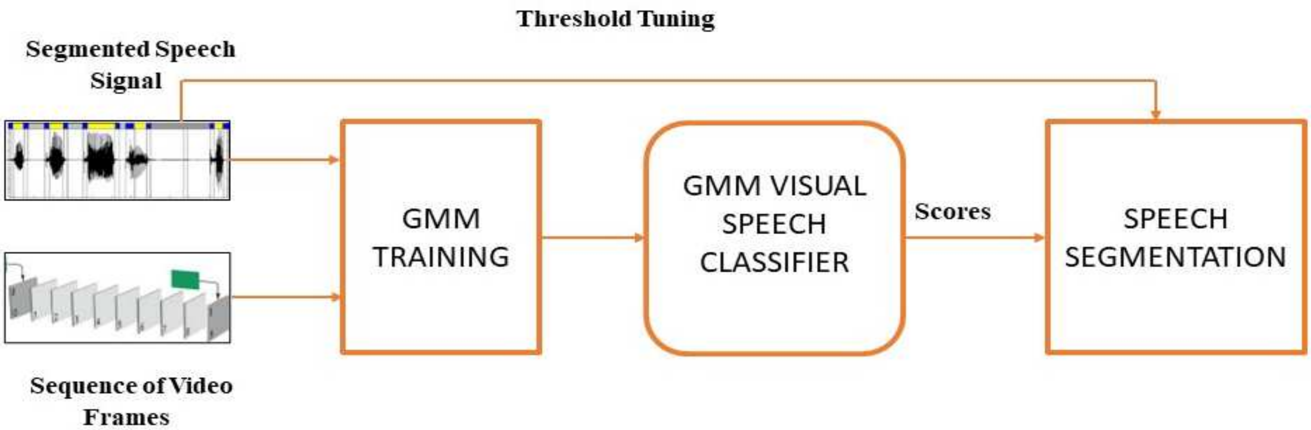

Fig.2: GMM based Video Voice Detection

B. Energy threshold methods

VAD systems can work in either the time domain or the frequency domain. Time domain techniques often employ energy and zero crossing rate (ZCR) parameters, whereas frequency domain techniques employ spectrum data. The energies of a speech frame provide information about a frame's activity, and the energy threshold value is employed in decision making. The amplitude of a speech in a frame is a crucial variable for time domain VAD techniques to identify frames as voice-active or inactive. In voice-active parts, speech sounds are classified as voiced, unvoiced, or plosives based on their patterns of excitation, with voiced phonemes (vowels) having quasi-periodic pulses and unvoiced phonemes (consonants) having random pulses. Plosive

sounds are excited in the same way as unvoiced noises are. The magnitude of voiced phonemes is substantially greater than that of unvoiced and plosive phonemes. Though unvoiced and plosive sounds have smaller amplitudes than voiced sounds, they include crucial information for speech, particularly when determining the start and finish points of a speech. The ZCR increases dramatically for an unvoiced speech at the starting point or extremities of an articulation when energy gets near to silence energy. Alternatively, the energy of vocal discourse is significantly more than the energy of stillness. Authors in [16] examine an energy-based procedure and presented an adaptive threshold valued VAD technique. This approach does not employ ZCR and relies solely on STE estimates. Using an adaptive scaling parameter, the programme dynamically estimates the energy threshold value. In each frame, the noise power is estimated and utilised to adjust the threshold. The study evaluates an adaptive thresholding problem and proposes potential solutions. They employed a detector based on linear energy with a twofold threshold. This procedure updates energy for both spoken and unvoiced frames based on updated threshold values. If the energy is larger than the threshold, the algorithm treats the frame as voiced; contrary, it treats it as unvoiced. It attempts to avoid the problem of abrupt variations in thresholding values by using distinct thresholding values for speech and silence detection. Jing et al.[17] has proposed a total spectrum energy based voice activity detection method. The wider frequency band noise energy is eliminated from the shorter frequency band noisy speech spectrum. Furthermore, a moving average filter is employed for smoothing the energy waveform of the speech spectra. The suggested method for detecting vocal activity is durable and performs well for an extensive range of SNR values. The proposed approach is resistant to varying SNR levels. For voice detection, it employs the short-time time/frequency features of the conveyed signal. Generally, human speech pitch and harmonics are unaffected by noise. A short-time FFT can be used to retrieve the sum of the pitch and harmonic spectrum energy in each speech window frame. The human voice has higher vitality in the low frequency region. Figure 3 illustrates an energy-enabled speech detection system.

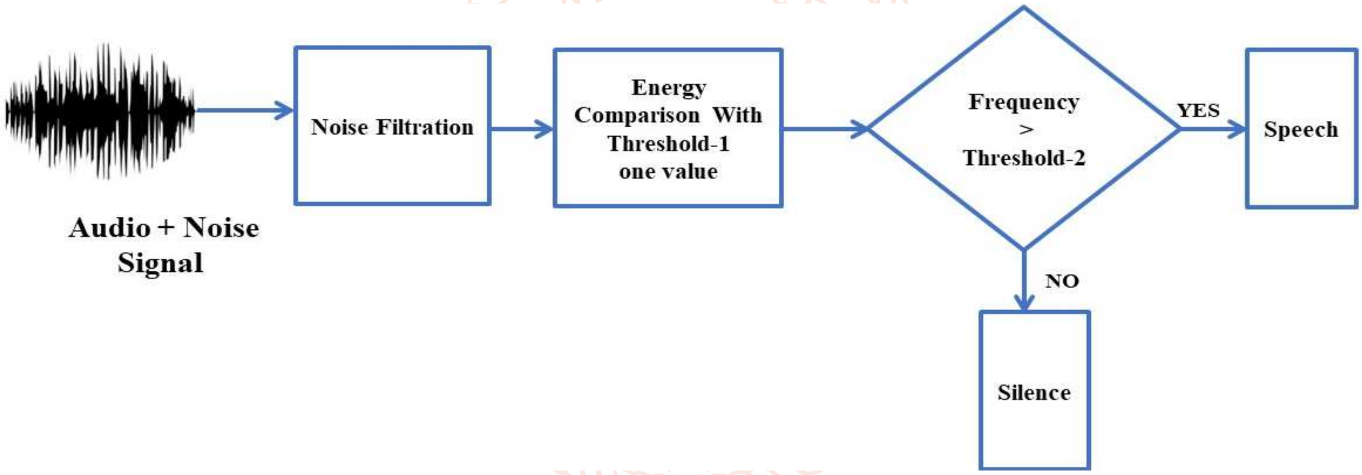

Fig. 3: Energy-enabled speech detection mechanism

C. Deep neural network (DNN) based methods

Speech recognition using neural networks is quite powerful. There are several networks for this reason. For speech recognition, RNN, LSTM, Deep Neural network, and composite HMM-LSTM are utilised. Authors in [18] have suggested a DNN based VAD method. They investigate contextual information at three levels using machine learning approaches. At the highest level, they used an MRS ensemble learning system, which is an array of ensemble predictors. Every predictor in a building block takes as input the combination of its lower building block's predictions and the enlargement of the raw acoustical feature by a given window. At the most basic level, they use the multi-resolution cochleagram characteristic, which combines contextual data by combining cochleagram features with different spectrotemporal resolutions. The experimental findings reveal that the MRS-based VAD surpasses other VADs significantly. Authors in [19] have proposed a DBN based VAD technique. It is effective at combining the benefits of several characteristics and provides cutting-edge functionality. The deep layers of the DBN-based VAD, however, do not appear to be superior to the shallower layers. To overcome this particular difficulty, they present a denoising-DNN based VAD in this research. In particular, they pre-train a deep neural network in an unsupervised denoising greedy layer-wise mode before fine-tuning the entire network in a supervised manner using the conventional back-propagation approach. During the pre-training phase, they use the noisy speech signals as the visible layer and attempt finding a novel characteristic that minimises the reconstruction cross-entropy loss between the noisy and clean speech signals. Investigational findings show that the suggested DDNN-enabled VAD not only surpasses the DBN-based VAD but additionally demonstrates that deep layers surpass shallower levels.

Authors in [20] described a different single-channel technique for VAD. They employ a Convolutional Neural Network (CNN) that makes use of spatial information. After extracting a frame-wise embedding series from the noisy input spectrum, a Self-Attention (SA) Encoder is used to find context-relevant data from the embedding series. Unlike prior research that used each frame (with context frames) independently, this system can handle the full signal at once, allowing for a large receptive field. They show that combining CNN and SA designs surpass techniques that only use CNN and SA. The authors in [21] introduced an approach for detecting vocal activity (VAD) based on CNN. The suggested method detects frames of voice in a specific audio source using the audio spectrogram raw image. The spectrogram is divided into frames that are categorised based on the existence or lack of a voice. The suggested technique outperformed state-of-the-art VAD algorithms in terms of accuracy under various noise situations. When evaluated on the QUT-NOISE-TIMIT database, it surpasses the top VAD systems with a considerable improvement in HTER. This method shows that utilising CNN on audio spectrogram images can be an effective approach to detecting voice even in highly distorted audio inputs. Fig. 4 shows the mechanism of a neural network-based VAD system.

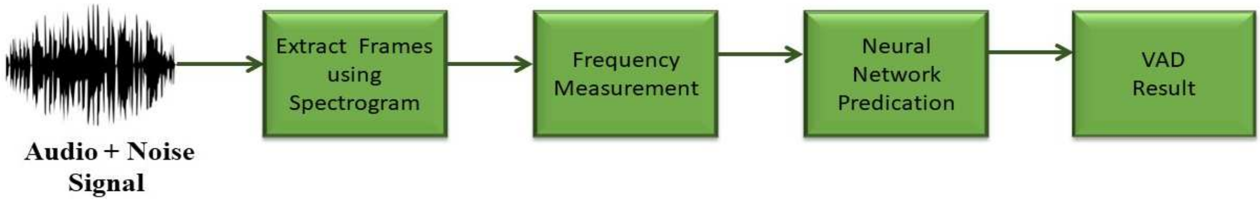

Fig. 4: Neural-Network-Enabled VAD System

D. Maximum margin based unsupervised VAD

For statistical voice activity detection (VAD), the authors [22] offered a novel robust feature and an unsupervised learning methodology. As an unsupervised classifier, maximum margin clustering (MMC) may boost the endurance of support vector machine (SVM)-based VAD while necessitating no data labelling for model training. The numerous observation compound feature (MO-CF) is introduced in the MMC framework to enhance correctness. Multiple observational signal-to-noise ratio (MO-SNR) and numerous observation maximum probability (MO-MP) are the two subfeatures of MO-CF. The contributions of the two subfeatures are equalised by a factor set to produce the highest area under the ROC curve (AUC) of performance. In studies covering several noisy settings with low SNRs, the suggested approach outperforms 7 commonly employed VAD algorithms.

CONCLUSION

VAD is a technique for detecting the existence or lack of human speech. A procedure can be started as a result of the detection. VAD has been used in speech-controlled programmes and gadgets such as cellphones that can be managed via voice commands. This paper studies various methods for VAD depending on various features. We classify the characteristics based on the attributes that are used, such as power, harmonicity, or modulation, and assess the performance of particular specialised elements. We have provided an overview of established ways to categorise characteristics based on the speech attributes that are used. Our analyses revealed that feature performance varies even when an identical speech characteristic is analysed.

REFERENCES

[1] M. Kotti, E. Benetos, C. Kotropoulos, I. Pitas, A neural network approach to audio-assisted movie dialogue detection, Neurocomputing 71 (1–3) (2007), pp.157–166.

[2] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Processing Letters 6 (1) (1999), pp. 1–3.

[3] WebRTC VAD, url:https://webrtc.org/.

[4] L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques, arXiv preprint arXiv:1003.4083.

[5] A. Misra, Speech/nonspeech segmentation in web videos, in: in: Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[6] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, H. Li, Voice activity detection using mfcc features and support vector machine, in: Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia, Vol. 2, 2007, pp. 556–561.

[7] C.E. Galván-Tejada, J.I. Galván-Tejada, J.M. Celaya-Padilla, J.R. Delgado-Contreras, R. Magallanes-Quintanar, M.L. Martinez-Fierro, I. Garza-Veloz, Y. López-Hernández, H. Gamboa-Rosales, An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and

neural networks, Mobile Information Systems (2016).

[8] N. Ryant, M. Liberman, J. Yuan, Speech activity detection on youtube using deep neural networks., in: INTERSPEECH, Lyon, France, 2013, pp. 728–731.

[9] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely`, P. Mateˇjka, Developing a speech activity detection system for the darpa rats program, in: Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[10] S. Tong, H. Gu, K. Yu, A comparative study of robustness of deep learning approaches for vad, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5695–5699.

[11] L. Mateju, P. Cerva, J. Zdánsky´ , J. Málek, Speech activity detection in online broadcast transcription using deep neural networks and weighted finite state transducers, in: 2017 IEEE International Conference on Acoustics, Speech andSignal Processing, ICASSP 2017, New Orleans, LA, USA, March 5–9, 2017, 2017, pp. 5460–5464.

[12] I. Jang, C. Ahn, J. Seo, Y. Jang, Enhanced feature extraction for speech detection in media audio, in: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20–24, 2017, 2017, pp. 479–483.

[13] X. Zhang, D. Wang, Boosting contextual information for deep neural network based voice activity detection, IEEE ACM Trans, Audio Speech Lang. Process. 24(2) (2016), pp. 252–264.

[14] X. Wu, M. Zhu, R. Wu and X. Zhu, "A Self-adapting GMM based Voice Activity Detection," 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, 2018, pp. 1-5, doi: 10.1109/ICDSP.2018.8631856.

[15] N. Rajitha, D. Sridharan, S. Fookes, C. Lucey and Patrick, "Visual Voice Activity Detection Using Frontal versus Profile Views", In The International Conference on Digital Image Computing : Techniques and Applications (DICTA2011), 6-8 December 2011.

[16] K. Sakhnov, E. Verteletskaya and B. Simak, "Approach for Energy-Based Voice Detector with Adaptive Scaling Factor," IAENG International Journal of Computer Science, 36, 2009.

[17] J. Pang, "Spectrum energy based voice activity detection," 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2017, pp. 1-5.

[18] X. -L. Zhang and D. Wang, "Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 2, pp. 252-264, Feb. 2016, doi:10.1109/TASLP.2015.2505415.

[19] X. -L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 853-857, doi: 10.1109/ICASSP.2013.6637769.

[20] Amit Sofer and Shlomo E. Chazan, "CNN self-attention voice activity detector," arXiv:2203.02944v1 [cs.SD] , 6 Mar 2022.

[21] Silva, D.A., Stuchi, J.A., Violato, R.P.V., Cuozzo, L.G.D. (2017). Exploring Convolutional Neural Networks for Voice Activity Detection. In: Paradisi, A., Godoy Souza Mello, A., Lira Figueiredo, F., Carvalho Figueiredo, R. (eds) Cognitive Technologies. Telecommunications and Information Technology. Springer, Cham.

[22] J. Wu and X. -L. Zhang, "Maximum Margin Clustering Based Statistical VAD With Multiple Observation Compound Feature," in IEEE Signal Processing Letters, vol. 18, no. 5, pp. 283-286, May 2011, doi:10.1109/LSP.2011.2119482.